Loners stand out. Identification of anomalous subsequences based on group performance

Martha Tatusch[0000−0001−6302−6070], Gerhard Klassen[0000−0002−1458−6546], and Stefan Conrad[0000−0003−2788−3854]

Heinrich Heine University, Universitätsstr. 1, 40225 Düsseldorf, Germany
{tatusch,klassen,stefan.conrad}@hhu.de

Abstract. Time series analysis is a part of data mining and nowadays an important field of research due to the increasing amount of data that is recorded sequentially by various systems. Especially the identification of anomalous subsequences arouses great interest, since a manual search for errors or malfunctions is not possible in most cases. Often outliers are defined as points or sequences that deviate significantly from the course of one or multiple time series, yet there are also applications where the trend rather than the exact course of time series is relevant. In that case, there is an approach of clustering the time series per time point and analyzing their cluster transitions over time. Sequences that change their cluster members suddenly or often, indicate an anomaly.
In 2019, a novel approach for the detection of these transition-based outliers was introduced [19]. Now, we present an algorithm called DACT (**D**etecting **A**nomalies based on **C**luster **T**ransitions) that is able to identify outlier sequences of the same type. It is a simple approach that stands out due to different results, although a similar type of anomalies is targeted. In the evaluation, we examine and discuss the differences. Our experiments show, that the results are competitive and reasonable.

Keywords: Outlier Detection · Time Series Analysis · Clustering.

1 Motivation

Due to the increasing popularity of digital systems such as social platforms, online shops or simple database applications in various industries, data analysis is of steadily growing importance. The analysis of sequential data forms an important part of this field of research and is known as *time series analysis*. There are several applications which consider either single or multiple time series whereby these can be univariate or multivariate. In this work, we focus on multiple multivariate time series and the behavior of subsequences with regard to their peers. There are many applications where these conditions apply. For example, when investigating a drug's tolerance on humans, one time series per patient can be extracted whereby various features per timestamp are recorded. In our approach, we examine the trend of groups of time series rather than the exact course, as it is not relevant in many applications. To do so, it is necessary to previously cluster the data for each point in time. Regarding the drug tolerance behavior, the patients may be grouped by their state of health. Since every human body is unique, these clusters may change over time. Some of these changes are normal, but if a patient shows any irregularity, action must be taken. In order to detect such irregularities automatically, we introduce DACT (**D**etecting **A**nomalies based on **C**luster **T**ransitions), an anomaly detection algorithm for transition-based outliers. To the best of our knowledge, the first approach regarding this type of outliers was published in 2019 [19]. Hence, in the following we will compare DACT with it.

2 Foundation

In order to provide a good basis for the comparison of the two methods, the same definitions as given in [19] are used in this work.

Definition 1 (Time Series). *A multivariate time series $T = o_{t_1}, ..., o_{t_n}$ is an ordered set of n real valued data points of arbitrary dimension. The data points are chronologically ordered by their time of recording.*

Definition 2 (Data Set). *A data set $D = T_1, ..., T_m$ is a set of m time series of same length and equivalent points in time. The set of data points of all time series at a timestamp t_i is denoted as O_{t_i}.*

Definition 3 (Subsequence). *A subsequence $T_{t_i,t_j,l} = o_{t_i,l}, ..., o_{t_j,l}$ with $j > i$ is an ordered set of successive real valued data points beginning at time t_i and ending at t_j from time series T_l.*

Definition 4 (Cluster). *A cluster $C_{t_i,j} \subseteq O_{t_i}$ at time t_i, with $j \in \{1, ..., q\}$ being a unique identifier (e.g. counter), is a set of similar data points, identified by a cluster algorithm or human.*

Definition 5 (Cluster Member). *A data point $o_{t_i,l}$ from time series T_l at time t_i, that is assigned to a cluster $C_{t_i,j}$ is called a member of cluster $C_{t_i,j}$.*

Definition 6 (Noise). *A data point $o_{t_i,l}$ from time series T_l at time t_i is considered as noise, if it is not assigned to any cluster.*

Definition 7 (Clustering). *A clustering is the overall result of a clustering algorithm or the set of all clusters annotated by a human for all timestamps. In concrete it is the set $\zeta = \{C_{t_1,1}, ..., C_{t_n,q}\} \cup Noise$.*

3 Related Work

There are various approaches for identifying irregularities in time series. In some applications, the detection of single anomalous data points is of interest. This problem is for example addressed by prediction-based algorithms like auto-regressive-moving-average (ARMA) models [2, 6, 15]. In other cases, the identification of so called *changing points* [7, 13], which indicate a change of the previous course, are relevant. Although these techniques perform very well in most cases, they can not be used for our purpose. First, in contrast to DACT, they target single data points, not subsequences. Second, they lack the correlation of one time series to others. There are also other algorithms for the detection of outliers, which decompose the time series with techniques like STL [4] before analyzing them. However, these methods only work if the considered time series can be actually decomposed. In many applications, this is not the case. When regarding anomalous subsequences, there are various works using dynamic time warping (DTW) [17] for the comparison of time series or neural networks [3, 10, 16]. Another approach is the detection of the most unusual subsequences (discords) using a symbolic aggregation of a time series [8, 12, 9]. Even though these methods are aiming at subsequences, they only consider single time series and therefore can not be used in our case.

The most recent works for the detection of outlier subsequences in multiple time series use Probabilistic Suffix Trees (PST) [18] or Random Block Coordinate Descents (RBCD) [21] regarding the deviation of one time series to the others. In contrast to our approach, the behavior of a time series with regard to its peers is not analyzed here. We accomplish this analysis by clustering the time series data per timestamp and investigating a time series' transitions between clusters. Such an approach was already presented in 2019 [19]. However, the procedure has some particularities that might be unfavorable depending on the application. For example, the procedure in [19] only penalizes splits of a time series from a cluster, whereas merges of smaller clusters into larger ones do not have a negative influence on the outlier score of the sequences involved. In this paper we introduce a simple approach which resolves these difficulties.

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

4 Model Description

After the time series data has been clustered per timestamp using an arbitrary clustering algorithm like DBSCAN [5] or k-means [14], DACT can be applied. In short, the procedure is based on the analysis of the average number of points in time that a time series migrates with its peers, which indicates a subsequence's stability over time. The longer a sequence moves with its cluster members over time, the more stable it is. For the following presentation of the components of DACT we first introduce the cluster identity function *cid* of a data point $o_{t_i,l}$, which returns the cluster of the time series l at the considered timestamp t_i:

$$cid(o_{t_i,l}) = \begin{cases} \emptyset & \text{if } o_{t_i,l} \text{ is not assigned to any cluster} \\ C_{t_i,a} & \text{else} \end{cases}$$

Now, we can calculate the number of time points in which two subsequences $T_{t_i,t_j,l}$ and $T_{t_i,t_j,x}$ share the same cluster. We call it the shared time points count *stc*:

$$stc(T_{t_i,t_j,l}, T_{t_i,t_j,x}) = |\{t_k | cid(o_{t_k,x}) = cid(o_{t_k,l}) \ \wedge \ t_k \in [t_i, t_j]\}|$$

with $x \neq l$. In order to get the average number of time points a time series $T_{t_i,t_j,l}$ moves with its cluster members, we need to compute the number of peers of the time series during the considered time period. It describes the amount of distinct time series that are at least once assigned to the same cluster as T_l during the period. It can be calculated by the peer count *pc*:

$$pc(T_{t_i,t_j,l}) = |\{T_x \ | \exists t_k \in [t_i, t_j] : cid(o_{t_k,x}) = cid(o_{t_k,l})\}|$$

with $x \neq l$. We can now express the over-time stability OTS of a subsequence $T_{t_i,t_j,l}$ by

$$OTS(T_{t_i,t_j,l}) = \frac{\sum_{p=1}^{m} stc(T_{t_i,t_j,l}, T_{t_i,t_j,p})}{pc(T_{t_i,t_j,l}) \cdot k}$$

with k being the number of timestamps where T_l holds data. In order to detect anomalies in time series, this score needs to be included in an outlier score, which indicates whether a subsequence is conspicuous or not. In the following we propose two concepts for building the outlier score. Since we believe, that this score is dependent on the behavior of a subsequence's peers (an unstable sequence is not as conspicuous regarding an unstable cluster as it is in a stable one), both variants focus on the scores of the considered cluster. Before introducing these two concepts, we define the term *intuitive outlier*:

Definition 8 (Intuitive Outlier). *A sequence $T_{t_i,t_j,l}$ is called an intuitive outlier if its data points are marked as noise for every timestamp $t_k \in [t_i, t_j]$.*

This is necessary as the outlier score can only be calculated for subsequences whose data point at the last timestamp is assigned to a cluster. If it is not, it is not possible to determine a meaningful reference value.

4.1 Variant 1

The first approach focuses on the best stability score achieved in a cluster $C_{t_j,a}$ regarding a time period from t_i to t_j. Formally, it can be expressed by

$$best\_score(C_{t_j,a}, t_i) = max(\{OTS(T_{t_i,t_j,l}) \mid cid(o_{t_j,l}) = C_{t_j,a}\}).$$

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

It describes the highest score obtained by subsequences from t_i to t_j ending in cluster $C_{t_j,a}$. The outlier score DACT of a subsequence is then given by the deviation of its stability score from the best score:

$$DACT(T_{t_i,t_j,l}) = best\_score(cid(o_{t_j,l}), t_i) - OTS(T_{t_i,t_j,l}).$$

Obviously, the *best_score* represents the upper bound for the outlier score within a cluster for a given time period. This causes, that clusters containing stable subsequences are more sensitive to deviations than the ones containing less stable sequences. Finally, an outlier can be formally described using the outlier score.

Definition 9 (Outlier – Variant 1). *Given a threshold τ, a sequence $T_{t_i,t_j,l}$ is called an outlier if*

$$DACT(T_{t_i,t_j,l}) > \tau \ .$$

Since the best subsequence score of a cluster influences the highest possible outlier score, the threshold τ often has to be chosen rather central in the interval $[0, 1]$. Additionally, the best threshold differs for data sets with different distributions of the data points. The more scattered the data, the lower the threshold.

4.2 Variant 2

The second approach follows the statistical assumption that anomalies can be found with the help of their deviation from the standard deviation. For this, the mean of a cluster's stability scores regarding the start time t_i has to be determined first. Regarding a cluster $C_{t_j,a}$ for the time period from t_i to t_j, it is given by

$$\mu(C_{t_j,a}, t_i) = \frac{1}{|C_{t_j,a}|} \cdot \sum_{o_{t_j,l} \in C_{t_j,a}} OTS(T_{t_i,t_j,l}).$$

The standard deviation of a cluster's stability scores regarding the start time t_i can then be calculated by

$$\sigma(C_{t_j,a}, t_i) = \sqrt{\frac{1}{|C_{t_j,a}|} \cdot \sum_{o_{t_j,l} \in C_{t_j,a}} (\mu(C_{t_j,a}, t_i) - OTS(T_{t_i,t_j,l}))^2}.$$

In order to compare it later with the standard deviation, we formulate the outlier score sDACT of a subsequence $T_{t_i,t_j,l}$ as the absolute difference of its stability score and the mean of its last cluster:

$$sDACT(T_{t_i,t_j,l}) = |\mu(cid(o_{t_j,l}), t_i) - OTS(T_{t_i,t_j,l})|.$$

We call it sDACT in order to express, that the *statistical* variant is used. In the following, this score can be used to detect outliers by inspecting the deviation of it from the standard deviation. With the help of a factor ρ it can be formally described.

Definition 10 (Outlier – Variant 2). *Given a threshold ρ, a sequence $T_{t_i,t_j,l}$ is called an outlier if*

$$sDACT(T_{t_i,t_j,l}) > \rho \cdot \sigma(cid(o_{t_j,l}), t_i) \ .$$

Again, the outlier score is highly dependent on the performance of the considered cluster's members. Since the standard deviation is considered, the outlier score is even less sensitive to deviations, especially in the case of a rather unstable cluster. Therefore in most cases the default value of $\rho = 3$ will probably be to high in order to detect inconsistencies. In our method, frequently a value of around $\rho \approx 2$ is recommended. This factor naturally is also dependent on the distribution of the data.

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

5 Experiments

Following, experiments on a synthetic and a real world data set are discussed to evaluate the performance of the presented methods. In order to simplify referencing the approaches we will name them as follows:

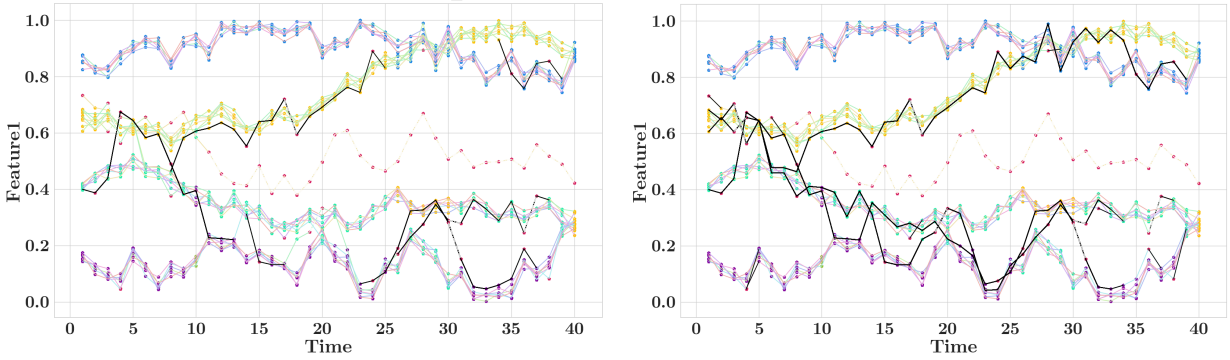
- *referred method* – describes the approach from [19].
- *DACT* – stands for the presented method using variant 1 for the detection of outliers.
- *sDACT* – represents the approach using variant 2.

5.1 Artificially Generated Data Set

The first considered data set was artificially generated and contains 28 univariate time series (TS) with 40 timestamps. Initially four groups of TS were randomly generated. Afterwards, three targeted and one completely random outlier sequence were inserted. All data points of the completely random outlier TS were chosen randomly, whereby the distance between two consecutive points was set to not being greater than 0.1. The remaining outlier sequences were generated so that their data points were always located near to a cluster's centroid. An outlier sequence could change its cluster at the earliest if it was located for at least 5 time points in a cluster.

The experiment was performed with DACT and the referred method. In order to get comparable results, the same parameter settings for both approaches were chosen. For the clustering DBSCAN [5] was used with $\epsilon = 0.025$ and $minPts = 3$. The threshold τ was set to 0.55. Figure 1 shows the detected anomalies by DACT and the referred method. The colored dots represent cluster belongings whereby red dots indicate noise. The detected outlier sequences are illustrated as and intuitive outliers as dashed lines.

Both methods managed to detect the completely random as well as parts of the three targeted outliers. The referred method, however, marked a lot more parts as outliers than DACT. Regarding the uppermost outlier sequence from time point 10 to 39, there is a difference between both methods between time 25 and 34. DACT did not mark this part of the TS as an outlier although the referred method did. This can be explained by the fact, that the TS moves stably with most of its cluster members in this period. The merge



(a) DACT (b) referred method

Fig. 1: Detected outliers on the generated data set with $\boldsymbol{\tau = 0.55}$, $\boldsymbol{minPts = 3}$ and $\boldsymbol{\epsilon = 0.025}$.

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

of the two upper clusters causes lower stability scores, but since the size of both clusters is approximately the same, all cluster members are affected equally. The same applies to the split.

Considering the second lowest outlier sequence between timestamp 30 and 38, it is the other way around. While DACT marks the sequence as an outlier for the whole period, the referred method interprets the course between timestamp 34 and 36 as normal. On the one hand, this is caused by the decrease of the stability scores in the second lowest cluster. As there were merges and splits in the history of the cluster, all scores were negatively affected. On the other hand, there are only few members in the considered cluster and another sequence is marked as noise at time point 32, too. Between timestamp 34 and 36 the considered time series behaves stable, so that it does not stand out in contrast to its cluster members, regarding this short period. In contrast to that, DACT is more sensitive concerning short term changes, if only few time series are considered.

5.2 GlobalEconomy Data Set

The second data set is provided by the website theglobaleconomy.com [1]. It consists of over 300 indicators for different features of 200 countries for more than 60 years. For the experiments, we considered 20 different countries and two features (namely the education spendings and the unemployment rate) within the period from 2010 to 2015 to enable a manageable illustration. Since the database is not complete for all country-year combinations, the amount of countries per timestamp may vary.

The experiment was run with all three methods using DBSCAN with $\epsilon = 0.19$ and $minPts = 2$. Since the underlying clustering for all three approaches is the same, it is illustrated separately in Figure 2. Different colors represent different cluster belongings and noise data points are marked red. The resulting outlier sequences are listed in Table 1. The list was shortened so that in case of overlaps only the longest detected subsequence of a country is included per method. This time, the threshold parameters τ and ρ were chosen for all methods separately, as the first experiment showed that the same parameter setting led to considerably more outlier sequences with the referred method than with DACT. An individual parameter choice might therefore be appropriate.

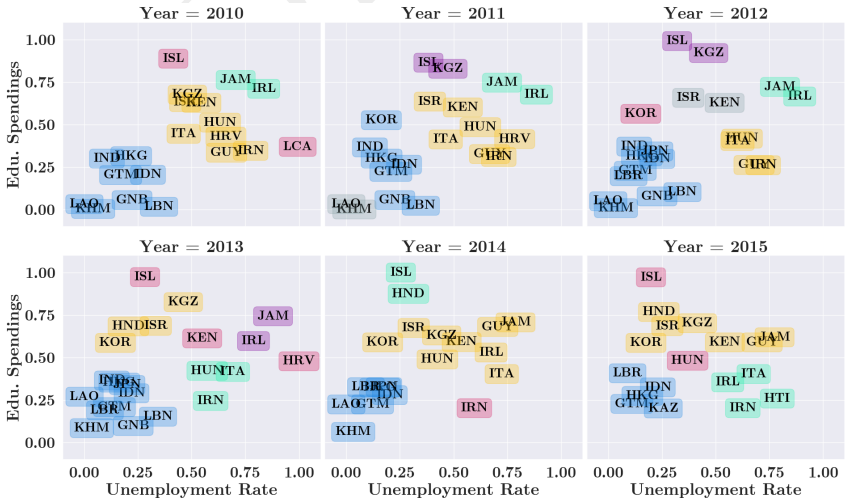


Fig. 2: Resulted clustering by DBSCAN with $\boldsymbol{minPts = 2}$ and $\boldsymbol{\epsilon = 0.19}$ on the GlobalEconomy data set.

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

It can be seen, that sDACT produces significantly less outlier sequences than DACT and the referred method. While those approaches detect both five anomalous subsequences, sDACT only finds two. This can be explained by the fact, that there are many clusters with only few cluster members. In addition, there are only a few TS, that are very stable over time. This causes, that the mean stability score per cluster is rather low. In order to stand out, a sequence needs therefore a very bad stability score. This only happens in two cases. First, Honduras (HND) does badly from 2014 to 2015, as it moves away from its only cluster member Iceland (ISL) and merges into a large cluster. The second case is Kenya (KEN) from 2013 to 2014, where it turns from noise to a large cluster's member. While the first anomaly sounds reasonable, the second one appears rather groundless, depending on the context. In contrast to DACT, which only found the first and not the second discussed outlier sequence, the referred method had exactly the opposite result. In fact, the only anomaly DACT and the referred method share, is the subsequence of Korea (KOR) from 2011 to 2014. This result is desired, since KOR changes its cluster members at every timestamp in this period.

Country	Start	End	DACT	sDACT	referred
GUY	2012	2015	–	–	x
HND	2013	2015	x	–	–
HND	2014	2015	x	x	–
IRL	2010	2014	x	–	–
JAM	2010	2014	x	–	–
KEN	2010	2015	–	–	x
KEN	2013	2014	–	x	x
KGZ	2010	2014	–	–	x
KOR	2011	2014	x	–	x

Table 1: Resulting outlier sequences by DACT ($\boldsymbol{\tau = 0.3}$), sDACT ($\boldsymbol{\rho = 2}$) and the referred method ($\boldsymbol{\tau = 0.35}$) on the GlobalEconomy data set.

The outlier sequences IRL and JAM show DACT's sensitivity regarding small clusters merging into large ones. Although those two countries stay stably together from 2010 to 2014, even when merging into the larger cluster, both are detected as outlier sequences. The referred method does not detect those sequences, because it does not penalize merges of clusters. However, although KEN stays with many cluster members over time, it is marked as outlier from 2010 to 2015. This is caused by the split from its cluster in 2012 and 2013. Another outlier detected by the referred method is Guyana (GUY) from 2012 to 2015. In 2013, the data is missing and this is the crucial point. In 2012 GUY is grouped with Hungary (HUN), Italy (ITA) and Iran (IRN). The merge into a larger cluster in 2014 is not penalized, but the following split from HUN, ITA and IRN in 2015 has a very negative effect on the stability, though.

6 Conclusion

In this paper, we introduced two approaches of finding transition-based outliers in time series databases. We examined the differences of the results and evaluated our methods against their competitor from [19], which targets the same problem definition. The results showed that both approaches find reasonable outliers, thus they differ in some characteristics. While the referred method does not penalize merges of clusters but only splits, DACT and sDACT treat both cases the same way. Furthermore, DACT is more sensitive regarding short term changes in small data sets. These differences lead to slightly different results, whereby the methods agree in clear cases. Depending on the application, both approaches provide a benefit.

We are aware of some shortcomings in DACT, that provide incentives for future work. For example, the handling of noise data points from the clustering could be improved. Currently, all subsequences consisting exclusively of noise data points are marked as intuitive outliers. In some cases, this behavior may not be legitimate. Furthermore, DACT is reliant on the assumption, that the underlying clustering is reasonable. Apart from inventing an evaluation measure for over-time clusterings [11, 20] in order to support the user in finding the right parameter settings, a new clustering algorithm tailored to the intention of an over-time clustering with temporal linkage would be useful.

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28

References

1. Global economy, world economy, https://www.theglobaleconomy.com/.
2. Ahmar, A.S., Guritno, S., Abdurakhman, Rahman, A., Awi, Alimuddin, Minggi, I., Tiro, M.A., Aidid, M.K., Annas, S., Sutiksno, D.U., Ahmar, D.S., Ahmar, K.H., Ahmar, A.A., Zaki, A., Abdullah, D., Rahim, R., Nurdiyanto, H., Hidayat, R., Napitupulu, D., Simarmata, J., Kurniasih, N., Abdillah, L.A., Pranolo, A., Haviluddin, Albra, W., Arifin, A.N.M.: Modeling data containing outliers using ARIMA additive outlier (ARIMA-AO). Journal of Physics: Conference Series **954** (2018)
3. Chambon, S., Thorey, V., Arnal, P.J., Mignot, E., Gramfort, A.: A deep learning architecture to detect events in eeg signals during sleep. In: 28th Int. Workshop on Machine Learning for Signal Processing. pp. 1–6 (2018)
4. Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.: Stl: A seasonal-trend decomposition procedure based on loess (with discussion). Journal of Official Statistics **6**, 3–73 (1990)
5. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. pp. 226–231 (1996)
6. Hill, D.J., Minsker, B.S.: Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environ. Model. Softw. **25**(9), 1014–1022 (2010)
7. Kawahara, Y., Sugiyama, M.: Change-point detection in time-series data by direct density-ratio estimation. In: Proceedings of the 2009 SIAM International Conference on Data Mining. pp. 389–400. SIAM (2009)
8. Keogh, E., Lin, J., Fu, A.: Hot sax: Efficiently finding the most unusual time series subsequence. In: Fifth IEEE International Conference on Data Mining (ICDM'05). p. 226–233 (2005)
9. Keogh, E., Lonardi, S., Chiu, B.Y.c.: Finding surprising patterns in a time series database in linear time and space. In: Proceedings of the 8th Int. Conference on Knowledge Discovery and Data Mining. pp. 550–556 (2002)
10. Kieu, T., Yang, B., Jensen, C.S.: Outlier detection for multidimensional time series using deep neural networks. In: 2018 19th IEEE Int. Conference on Mobile Data Management (MDM). pp. 125–134 (2018)
11. Klassen, G., Tatusch, M., Himmelspach, L., Conrad, S.: Fuzzy clustering stability evaluation of time series. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU (2020)
12. Lin, J., Keogh, E., Ada Fu, Van Herle, H.: Approximations to magic: finding unusual medical time series. In: 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05). pp. 329–334 (2005)
13. Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. Neural Netw. **43**, 72–83 (2013)
14. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297 (1967)
15. Munir, M., Siddiqui, S.A., Chattha, M.A., Dengel, A., Ahmed, S.: Fusead: Unsupervised anomaly detection in streaming sensors data by fusing statistical and deep learning models. Sensors **19**(11) (2019)
16. Munir, M., Siddiqui, S.A., Dengel, A., Ahmed, S.: Deepant: A deep learning approach for unsupervised anomaly detection in time series. IEEE Access **7**, 1991–2005 (2018)
17. Salvador, S., Chan, P.: Toward accurate dynamic time warping in linear time and space. Intell. Data Anal. **11**(5), 561–580 (2007)
18. Sun, P., Chawla, S., Arunasalam, B.: Mining for outliers in sequential databases. In: ICDM, 2006. pp. 94–106
19. Tatusch, M., Klassen, G., Bravidor, M., Conrad, S.: Show me your friends and i'll tell you who you are. finding anomalous time series by conspicuous cluster transitions. In: Data Mining. AusDM 2019. Communications in Computer and Information Science. vol. 1127, pp. 91–103 (2019)
20. Tatusch, M., Klassen, G., Bravidor, M., Conrad, S.: How is your team spirit? cluster over-time stability evaluation. In: Machine Learning and Data Mining in Pattern Recognition, MLDM (2020)
21. Zhou, Y., Zou, H., Arghandeh, R., Gu, W., Spanos, C.J.: Non-parametric outliers detection in multiple time series a case study: Power grid data analysis. In: AAAI (2018)

The final authenticated publication is available online at https://doi.org/10.1007/978-3-030-65390-3_28